Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Instance Specific Metric Subspace Learning: A Bayesian Approach*

Han-Jia Ye and **De-Chuan Zhan** and **Yuan Jiang**
National Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, 210023, China
{yehj, zhandc, jiangy}@lamda.nju.edu.cn

Abstract

Instead of using a uniform metric, *instance specific distance* learning methods assign multiple metrics for different localities, which take data heterogeneity into consideration. Therefore, they may improve the performance of distance based classifiers, e.g., kNN. Existing methods obtain multiple metrics of test data by either transductively assigning metrics for unlabeled instances or designing distance functions manually, which are with limited generalization ability. In this paper, we propose ISMETS (Instance Specific METric Subspace) framework which can *automatically* span the whole metric space in a generative manner and is able to *inductively* learn a specific metric subspace for each instance via inferring the expectation over the metric bases in a Bayesian manner. The whole framework can be solved with Variational Bayes (VB). Experiment on synthetic data shows that the learned results are with good *interpretability*. Moreover, comprehensive results on real world datasets validate the *effectiveness* and *robustness* of ISMETS.

Introduction

Many classifiers depend significantly on distances between examples, e.g., kNN and Gaussian kernel methods. Distance metric learning approaches are typically investigated with the purpose of learning a good Mahalanobis distance metric to pull similar instances together and push different ones away. With the learned metric, classifiers can make better predictions on test data. Most distance metric learning methods, such as LMNN (Weinberger, Blitzer, and Saul 2006) and ITML (Davis et al. 2007), focus on learning a uniform metric to measure distance between all pairs of instances.

However, in some applications, data heterogeneity need to be taken into consideration. Thus instead of learning a uniform distance metric for some concerned tasks, it is more reasonable for each instance to claim its own *instance specific distance* so that it can measure the distance to others from its own perspective. Although some instance specific distance learning methods have achieved advantages, they predict unlabeled instances transductively (Zhan et al. 2009), or model the instance specific distance for test data manually (Frome, Singer, and Malik 2007). Transductive approaches lack flexibility on predicting unlabeled data, i.e., unlabeled data should be included during the training phase, which is often violated in real cases. Manually designed distance functions make too much model assumptions which increase the risk of instability. In particular, Shi, Bellet, and Sha (2014) utilize fixed bases to form metrics by linear combination, and obtain the combination coefficients by optimizing a parametric model with arbitrary pre-defined functions. Both manually chosen bases and pre-defined functions may result in unstable performance.

In this paper, we focus on representing the instance specific distance by a generative model, where all parameters of *instance specific distance* can be learned/inferred automatically. We assume each instance $\mathbf{x}$, either labeled or unlabeled, has its own instance specific metric $M_{\mathbf{x}}$ for calculating distance to others. Each $M_{\mathbf{x}}$ is constructed within a subspace of metric space $\mathcal{M}$, i.e., each metric for a particular instance is within its own subspace spanned by only a portion of metric bases. In order to represent the instance specific metric subspace, we need to learn global bases $M = \{M_k\}_{k=1}^K$ of metric space, and span the metric subspace in a K-simplex for each instance. Therefore, we name the proposed model ISMETS (Instance Specific METric Subspace learning). In ISMETS, we embed the bases of metric space $\mathcal{M}$ into a generative process to *learn* the bases and metric subspace simultaneously in a Bayesian manner. We introduce latent allocation variables combining the bases to form the metric subspace. Besides, they can also bridge the latent information with a linear classifier, and therefore, incorporate the model with side information. The whole model is trained by variational Bayes and the posterior of metric for a specified instance can be inductively inferred in an effective and robust way. To the best of our knowledge, we are the first to learn the instance specific metric subspace with Bayesian model. In summary, we claim our method has two main advantages:

- ISMETS is a Bayesian generative model. It provides robust specific metric subspace for each instance by learning bases and combination coefficients at the same time;
- ISMETS can handle both labeled and unlabeled data, and can inductively infer instance specific metric for unseen test data as well.

*This work was supported by NSFC (61273301, 61321491) and CCF-Tencent Open Research Fund (RAGR20150117).
Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The rest of this paper starts with a review of related work. Then the ISMETS framework is presented in detail, which is followed by the experiments and conclusion.

Related Work

The basic idea of distance metric learning is to find a distance metric such that the distance between data points in the same class is smaller than that from different ones. Different methods use different criteria to achieve this goal, such as constrained convex programming (Xing et al. 2003), max-margin nearest neighbor (Weinberger, Blitzer, and Saul 2006), information theoretical approaches based on Bregman optimization (Davis et al. 2007). Generalization performance of metric learning is analyzed in (Jin, Wang, and Zhou 2010). Comprehensive reviews of metric learning can be found in (Kulis 2012; Bellet, Habrard, and Sebban 2013).

In recent years, instance specific distance metric learning becomes attractive, which adapts well to complex data patterns. Instead of learning a uniform metric for all instances, this kind of approach assumes that each instance should have its own metric for measuring the distance to others. Zhan et al. (2009) expand the idea of (Frome, Singer, and Malik 2007) to learn a specified metric for each instance by metric propagation technique. However, it can only deal with unlabeled data transductively. Weinberger and Saul (2009) use local information to train multiple metrics for instances in different data clusters. Shi, Bellet, and Sha (2014) propose a sparse compositional method, which assumes metrics are the combination of pre-defined bases and metrics for test data are generalized with parametric functions. Either predicting test data transductively or formulating the metric with arbitrary pre-defined distance functions might limit the generalization ability, and may eventually degenerate the classification performance, since these strategies lack flexibilities to handle polytropic data very well.

There are few Bayesian approaches for metric learning. Yang, Jin, and Sukthankar (2007) propose an active metric learning approach with probability model yet neither considers local structures nor learns with Bayesian inference. MMDML in (Babagholami-Mohamadabadi et al. 2014), only uses Bayesian method for classification in multimodal scenario while no metric is directly incorporated.

Different from these existing approaches, ISMETS is designed to provide the instance specific metric by *learning* metric bases and corresponding combination coefficients automatically. Within a unified Bayesian framework, ISMETS can learn the metric bases to form the metric subspace for each instance both transductively and inductively.

Our Proposed Approach

In this section, we first describe our Instance Specific METric Subspace (ISMETS) model in detail, and then put forward the variational Bayes training process together with the whole inductive inference process. Without loss of generality, for a C-class problem, we suppose there are N training instances represented as $X = \{\mathbf{x}_n \in \mathbb{R}^d\}_{n=1}^N$. First N_1 of X are labeled examples with label $y_n \in \{1, \cdots, C\}$ for $\mathbf{x}_n$. The latter N_2 are unlabeled instances.

Instance Specific Metric Subspace Model

Common metric learning methods aim to learn a uniform metric shared by all instances to compute distances with others. Nevertheless, numerous real applications desire that each instance should have its own distance metric (Hu et al. 2015). In ISMETS, we treat the Instance Specific Metric (ISM) as a mapping from instances to metric, i.e., the distance metric for a specific instance $\mathbf{x}_n$ can be formulated as a function $g : \mathbf{x}_n \to \mathcal{S}_+^d$, which can be defined as $g(\mathbf{x}_n) = M_{\mathbf{x}_n}$ and $\mathcal{S}_+^d$ represents a $d \times d$ positive semidefinite matrix. The (squared) Mahalanobis distance from $\mathbf{x}_n$ to another instance $\mathbf{x}_m$ can be denoted as:

$$D(\mathbf{x}_n, \mathbf{x}_m) = (\mathbf{x}_n - \mathbf{x}_m)^\top M_{\mathbf{x}_n} (\mathbf{x}_n - \mathbf{x}_m). \quad (1)$$

It is notable that the instance specific distance in Eq. 1 is non-symmetric and the metric used is based on the first instance in the pair. In particular, the ISM $M_{\mathbf{x}_n}$ is assumed to be sampled from a global metric space $\mathcal{M}$ which is spanned by metric bases $M = \{M_k \in \mathcal{S}_+^d\}_{k=1}^K$, following a latent allocation distribution $\mathbf{z}_n = [\mathbf{z}_{n1}, \mathbf{z}_{n2}, \cdots, \mathbf{z}_{nK}]$ (Blei, Ng, and Jordan 2003). We model the allocation variables $\mathbf{z}_n$ as a latent component of a mixture model. It is usually assumed that $\mathbf{z}_n \in \{0, 1\}^K$ and follows a multinomial distribution. $\mathbf{z}_{nk}$ is the kth element of $\mathbf{z}_n$ and $\sum_k \mathbf{z}_{nk} = 1$. In this paper, we claim that each ISM $M_{\mathbf{x}_n}$ can be constituted as

$$M_{\mathbf{x}_n} = \mathbb{E}_{\mathbf{z}_n}[M], \quad (2)$$

i.e., the $M_{\mathbf{x}_n}$ is an expectation over bases M on distribution $\mathbf{z}_n$. Since in allocation distributions all $\mathbf{z}_{nk} \in \mathbf{z}_n$ are nonnegative, $M_{\mathbf{x}_n}$ is ensured to be positive semidefinite as a valid metric. Different from the assumption made in (Shi, Bellet, and Sha 2014), ISMETS can learn the metric bases M and the compositional latent distribution $\mathbf{z}_n$ at the same time. This reduces the influence of the disagreement between assumptions and ground-truth distribution. In order to incorporate with supervision information, instead of employing pairwise or triplet constraints as general distance metric learning approaches, we directly embed a classifier related to $\mathbf{z}_n$ for simplification of modeling. This setting corresponds to the fully supervised metric learning in (Bellet, Habrard, and Sebban 2013).

We present the Probabilistic Graphical Model (PGM) of ISMETS in Fig. 1, where the left and right parts are generation process of two instances $\mathbf{x}_n$ and $\mathbf{x}_m$ respectively. Variables in the dotted box are only for labeled examples, i.e., for unlabeled ones there is no variable f related to $\mathbf{z}$. In particular, the whole model can be summarized as follows: instance $\mathbf{x}_n$ is drawn from a local distribution which is related to its own ISM subspace $\mathcal{M}_{\mathbf{x}_n} \subseteq \mathcal{M}$. This subspace shares global metric bases $M = \{M_k\}_{k=1}^K$ with others and is sampled according to a latent allocation distribution $\mathbf{z}_n$. Note that the locality is closely related to its neighbors $\mathbf{x}_m$ ($m \neq n$). Given bases M, it is assumed that the local property of instance $\mathbf{x}_n$ is coded by $\mathbf{z}_n$ which should be closely related to the class information. The variable $f_n \in \{-1, 1\}^C$, which indicates the label y_n of $\mathbf{x}_n$, is consequently sampled following the distribution of $\mathbf{z}_n$, and is controlled by classifier $\mathbf{w}$ and b. To simplify the discussion, we illustrate the model for binary classification (C = 2), and it can be easily extended to multi-class cases as in our experiments.

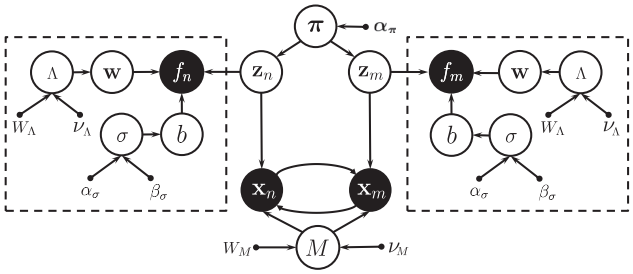


Figure 1: Probabilistic Graphical Model (PGM).

Key Components of The PGM

There are 3 key components in ISMETS, i.e., (1). the part of latent allocation for ISM subspace, which is about the choice of $\mathbf{z}_n$ and its conjugate prior; (2). the part of data generation from ISM subspace related to M, $\mathbf{z}_n$ and $\mathbf{x}_n$; (3). the part of label information injection depending on f_n, $\mathbf{z}_n$, $\mathbf{w}$ and b.

In ISMETS, each instance $\mathbf{x}_n$ has an ISM $M_{\mathbf{x}_n}$ from its own perspective, and then the distance from $\mathbf{x}_n$ to others can be calculated by Eq. 1. It is notable that each $M_{\mathbf{x}_n}$ is drawn from metric space $\mathcal{M}$ with respect to a distribution defined by latent allocation vector $\mathbf{z}_n$, and all metric subspaces from $\mathcal{M}$ share the metric bases M. $\mathbf{z}_{nk} = 1$ reveals the choice of metric base M_k for $\mathbf{x}_n$, so the overall conditional probability on the collection of all latent allocation variables $Z = \{\mathbf{z}_n\}_{n=1}^N$ is $Z|\boldsymbol{\pi} \sim \prod_{n=1}^N \prod_{k=1}^K \boldsymbol{\pi}_k^{\mathbf{z}_{nk}}$. Parameter $\boldsymbol{\pi} = [\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_K]$ and $\boldsymbol{\pi}_k$ is the kth element of $\boldsymbol{\pi}$. We further assume $\boldsymbol{\pi}$ with prior: $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_{\boldsymbol{\pi}})$, where $\mathrm{Dir}(\cdot)$ specifies a Dirichlet distribution with parameter $\boldsymbol{\pi}$ and hyper-parameter $\boldsymbol{\alpha}_{\boldsymbol{\pi}}$.

Inspired by the relationship between matter density and distances of molecules in physics, we assume the probability of sample points is determined by data density to a great extent, and the data density is defined by a distance-related function like Kernel Density Estimation (Chen, Chen, and Weinberger 2014). Therefore, in ISMETS, the conditional probability likelihood of instance $\mathbf{x}_n$ can be denoted as:

$$p(\mathbf{x}_n|M_{\mathbf{x}_n}, \mathbf{x}_{-n}) = \frac{1}{N-1} \sum\nolimits_{i \neq n} \mathcal{N}\left(\mathbf{x}_n|\mathbf{x}_i, M_{\mathbf{x}_n}^{-1}\right),$$

where $\mathbf{x}_{-n}$ denotes all instances except $\mathbf{x}_n$. The quadratic term in Gaussian distribution $\mathcal{N}$ uses the ISM $M_{\mathbf{x}_n}$ to determine the distance from $\mathbf{x}_n$ to $\mathbf{x}_i$. The impact of locality simultaneously depends on the location of $\mathbf{x}_i$ and $M_{\mathbf{x}_n}$. Combined with the i.i.d. property of X, we have $p(\mathbf{x}_n|M) \overset{\text{def}}{=} p(\mathbf{x}_n|M_{\mathbf{x}_n}, \mathbf{x}_{-n})$. The overall conditional probability $p(X|Z, M)$ on all training data, consequently, is:

$$p(X|Z, M) = \prod_{n=1}^N p(\mathbf{x}_n \mid M) = \prod_{n=1}^N \prod_{k=1}^K p(\mathbf{x}_n|M_{\mathbf{x}_n}, \mathbf{x}_{-n})^{\mathbf{z}_{nk}}. \tag{3}$$

Since the ISM $M_{\mathbf{x}_n}$ is sampled according to the latent allocation variable $\mathbf{z}_n$ in a metric subspace spanned by metric bases, the final representation of $M_{\mathbf{x}_n}$ can be turned into an expectation of metric bases over $\mathbf{z}_n$. Given an instance, labeled or unlabeled, we need to determine the ISM $M_{\mathbf{x}_n}$. To guarantee $M_{\mathbf{x}_n}$ a valid distance metric, we model M_k subject to a Wishart distribution, and thus the whole probability distribution of metric bases is:

$$M \sim \prod\nolimits_{k=1}^K \mathcal{W}(M_k|W_{M,k}, \nu_{M,k}), \tag{4}$$

where $W_{M,k}$ is a positive semidefinite matrix and $\nu_{M,k}$ is freedom degree of the distribution.

The latent allocation variable $\{\mathbf{z}_n\}_{n=1}^{N_1}$ of labeled data, which bridges M_k and $M_{\mathbf{x}_n}$, should concentrate on local properties of instance $\mathbf{x}_n$, and consequently can be employed in a classifier. It is noteworthy that although $\mathbf{z}_n \in \{0, 1\}^K$, the linear classifier really relies on the expectation of $\mathbf{z}_n$ in the inference procedure. Moreover, $\mathbf{z}_n$ can retain the classification information especially when K is relatively large. Therefore, we use a linear classifier $\mathbf{w}$ based on the latent variable $\mathbf{z}_n$ for labeled examples. In addition, a linear classifier $\mathbf{w}$ with respect to the latent allocation variable can be a nonlinear classifier from the aspect of the instance feature space. In general, $\mathbf{w}$ can be generated from a Gaussian distribution with its precision matrix Λ setting as a Wishart distribution prior. Similarly, the bias b of the classifier follows a Gaussian distribution with a gamma prior σ:

$$\begin{aligned} \mathbf{w}|\Lambda &\sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \Lambda^{-1}), & \Lambda &\sim \mathcal{W}(\Lambda|W_\Lambda, \nu_\Lambda), \\ b|\sigma &\sim \mathcal{N}(b, |0, \sigma^{-1}), & \sigma &\sim \mathcal{G}(\sigma|\alpha_\sigma, \beta_\sigma), \end{aligned}$$

where α_σ and β_σ are shape and scale parameters in gamma distribution respectively. Both $\mathbf{w}$ and b have zero mean for symmetry. Combining the classifier above, we can get the label distribution (Mcauliffe and Blei 2008):

$$p(F|\mathbf{w}, b, Z) = \prod\nolimits_{n=1}^{N_1} \mathcal{N}\left(f_n|\mathbf{w}^\top \mathbf{z}_n + b, 1\right),$$

where F is the collection of label indicator f_n. Intuitively, supervision part makes instances from the same class share similar latent allocation variable $\mathbf{z}$. This would yield similar ISMs for instances from the same class. This part also can inject supervision of labeled data and will guide the training phase. Therefore, ISMETS can work in a semi-supervised scenario. Moreover, by setting f_n as a label dummy vector and training C classifiers $\mathbf{w}_i, b_i, i = 1, \cdots, C$, one for each class (Xiang et al. 2012; Gönen and Margolin 2014), ISMETS can be easily extended to multi-class cases.

For further discussions, we denote the set of all latent variables as $\Omega = \{\boldsymbol{\pi}, Z, M, \mathbf{w}, \Lambda, b, \sigma\}$. The joint distribution of the PGM can be represented by:

$$\begin{aligned} p(X, F, \Omega) = p(\boldsymbol{\pi})p(Z|\boldsymbol{\pi})p(M)p(X|Z, M)p(\Lambda) \\ p(\mathbf{w}|\Lambda)p(\sigma)p(b|\sigma)p(F|\mathbf{w}, b, Z). \end{aligned}$$

Variational Inference on Proposed Model

The joint probability of ISMETS is complex and makes exact inference on the PGM intractable. Therefore, we use variational Bayes in model training phase. Concretely speaking, a variational distribution $q(\Omega)$ is used to approximate the true posterior distribution defined by $p(\Omega|X, F)$. With mean-field assumption (Jordan et al. 1999; Wainwright and Jordan 2008; Zhao et al. 2014), the approximation can be made as:

$$p(\Omega|X, F) \approx q(\Omega) = q(\boldsymbol{\pi})q(Z)q(M)q(\mathbf{w}, b)q(\Lambda)q(\sigma),$$

where the concrete forms of factors are:

$$q(M) = \prod_{k=1}^{K} q(M_k) = \prod_{k=1}^{K} \mathcal{W}(M_k|\hat{\nu}_k, \hat{W}_k),$$
$$q(Z) = \prod_{n=1}^{N}\prod_{k=1}^{K} \boldsymbol{\gamma}_{nk}^{\mathbf{z}_{nk}},$$
$$q(\mathbf{w}, b) = \mathcal{N}\left(\begin{bmatrix}\mathbf{w}\\ b\end{bmatrix} \mid \hat{\boldsymbol{\mu}}_{\mathbf{w}}, \hat{\Sigma}_{\mathbf{w}}\right),$$
$$q(\Lambda) = \mathcal{W}(\Lambda|\hat{W}_\Lambda, \hat{\nu}_\Lambda), \quad q(\sigma) = \mathcal{G}(\sigma|\hat{\alpha}_\sigma, \hat{\beta}_\sigma).$$

Note that comparing to the original PGM in Fig. 1, the variational version only makes some moderate independent assumptions on variables, and the approximation distributions remain the same distribution family as the original ones. E.g., $q(M)$ and $q(Z)$ are Wishart and multinomial distribution respectively with parameter $\hat{\nu}_k, \hat{W}_k$ and $\boldsymbol{\gamma}_{nk}$. These guarantee the consistency between the original PGM and the variational model as well as the convergence of inference (Watanabe and Watanabe 2006). In our implementations, ISMETS gets satisfactory results in 20 iterations. By minimizing the KL-divergence between the approximation $q(\Omega)$ and the true posterior $p(\Omega|X, F)$, update rules for distribution parameters in the supervision injection part can be summarized as:

$$\hat{\alpha}_\sigma = \alpha_\sigma + 0.5, \quad \hat{\beta}_\sigma^{-1} = \beta_\sigma^{-1} + 0.5 * \mathbb{E}[b^2],$$
$$\hat{\nu}_\Lambda = \nu_\Lambda + 1, \quad \hat{W}_\Lambda^{-1} = W_\Lambda^{-1} + \mathbb{E}[\mathbf{w}\mathbf{w}^\top],$$
$$\hat{\Sigma}_{\mathbf{w}}^{-1} = \begin{bmatrix}\mathbb{E}[\Lambda] + \sum_{n=1}^{N_1}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] & \sum_{n=1}^{N_1}\mathbb{E}[\mathbf{z}_n]\\ \sum_{n=1}^{N_1}\mathbb{E}[\mathbf{z}_n^\top] & \mathbb{E}[\sigma] + N_1\end{bmatrix},$$
$$\hat{\mu}_{\mathbf{w}} = \hat{\Sigma}_{\mathbf{w}} \cdot \begin{bmatrix}\sum_{n=1}^{N_1}\mathbb{E}[\mathbf{z}_n]f_n\\ \sum_{n=1}^{N_1} f_n\end{bmatrix},$$

where $\mathbb{E}[\cdot]$ denotes the expectation w.r.t. the approximate posterior $q(\Omega)$. It is difficult to get closed form update rules for parameters in $q(M)$, since inference on M is related to kernel density estimation in Eq. 3, and the expectation of the log-likelihood of this distribution contains a series of summations in the $\ln(\cdot)$ term. Hence we use Jensen's inequality to get a lower bound of $q(M)$ and finally obtain the following update rules:

$$\hat{\nu}_k = \nu_{M,k} + \sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_{nk}],$$

$$\hat{W}_k^{-1} = W_{M,k}^{-1} + \frac{1}{N-1}\sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_{nk}]\sum_{i\neq n}(\mathbf{x}_n - \mathbf{x}_i)(\mathbf{x}_n - \mathbf{x}_i)^\top,$$

where the degree of freedom is updated by the effective number of instance for each basis and the inverse of $\hat{W}_k$ is updated by the covariance matrix between instances, weighted by the latent allocation. Since the independency between updates of metric bases M, the whole training process can be conducted parallelly.

$q(\boldsymbol{\pi})$ is also a Dirichlet distribution, i.e., $q(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\pi}|\hat{\boldsymbol{\alpha}})$. $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^K$ with element updated by $\hat{\boldsymbol{\alpha}}_k = \boldsymbol{\alpha}_{\boldsymbol{\pi},k} + \sum_{n=1}^{N}\mathbb{E}[\mathbf{z}_{nk}]$, where $\boldsymbol{\alpha}_{\boldsymbol{\pi},k}$ is the kth element of $\boldsymbol{\alpha}_{\boldsymbol{\pi}}$.

For the latent variable Z, we should treat the labeled and unlabeled parts with different strategies, since Z not only depends on $\mathbf{x}$ and M but also is related to the supervision injection part for labeled data. For each instance $\mathbf{x}_n$, we first get the unnormalized term $\boldsymbol{\rho}_{nk}$ and then get the normalized parameter $\boldsymbol{\gamma}_{nk} = \frac{\boldsymbol{\rho}_{nk}}{\sum_j \boldsymbol{\rho}_{nj}}$. For unlabeled data, the expectation over logarithm in Eq. 3 makes difficulties on inference. Therefore, we approximate the results of $\boldsymbol{\rho}_{nk}$ with

$$\ln(\boldsymbol{\rho}_{nk}) = \mathbb{E}[\ln(\boldsymbol{\pi}_k)] + \mathbb{E}\left[\ln\left(\tfrac{1}{N-1}\textstyle\sum_{i\neq n}\mathcal{N}\left(\mathbf{x}_n|\mathbf{x}_i, M_k^{-1}\right)\right)\right]$$

$$\approx \mathbb{E}[\ln(\boldsymbol{\pi}_k)] + \ln\left(\frac{1}{N-1}\sum_{i\neq n}\exp\left(-\frac{d}{2}\ln 2\pi + U_i\right)\right),$$

$$U_i = \frac{1}{2}\mathbb{E}\left[\ln|M_k|\right] - \frac{1}{2}(\mathbf{x}_n - \mathbf{x}_i)^T\mathbb{E}[M_k](\mathbf{x}_n - \mathbf{x}_i).$$

In addition, the $\ln(\boldsymbol{\rho}_{nk})^{\mathrm{lb}}$ for labeled data should also absorb the label indicator information from classifier:

$$\ln(\boldsymbol{\rho}_{nk})^{\mathrm{lb}} = \ln(\boldsymbol{\rho}_{nk}) - \sum_{j\neq k} A_{kj}\mathbb{E}[\mathbf{z}_{nj}] - \frac{1}{2}A_{kk} + B_k^n,$$

where $A = \mathbb{E}[\mathbf{w}\mathbf{w}^\top] \in \mathbb{R}^{K\times K}$, $B^n = f_n\mathbb{E}[\mathbf{w}] \in \mathbb{R}^K$, and A_{kj} is the (k, j)-entry of matrix A, B_k^n is the kth element of B^n. A and B^n reveal the label information. It is obvious that the update of Z for the labeled examples is combined with $\mathbf{w}, b$ and label assignments.

Transductive and Inductive Prediction

ISMETS can handle both labeled and unlabeled data. So test instances can be regarded as unlabeled instances in the training procedure, i.e., metrics for test instances can be assigned transductively. In transductive configuration, each test instance $\mathbf{x}^*$ has its own allocation variable $\mathbf{z}^*$ and the ISM $M_{\mathbf{x}^*}$ for test instance equals to the expectation over learned metric bases, i.e., $M_{\mathbf{x}^*}$ is calculated according to Eq. 2.

In real world problems where generalization ability is desired, however, the model should output ISM for *unseen* test instances $\mathbf{x}^*$. Therefore, we need to figure out the posterior distribution of the ISM $M_{\mathbf{x}^*}$, by integrating out other latent variables. In detail, we can use the approximate distribution achieved in the training update process to help integrate the latent random variables. By substituting the approximations, it yields the posterior:

$$p(M_{\mathbf{x}^*}|\mathbf{x}^*, X, Y) = \sum_{k=1}^{K}\frac{\hat{\boldsymbol{\alpha}}_k}{\hat{\alpha}}\frac{1}{N}\sum_{n=1}^{N}\mathcal{W}(M_{k,n}|W_{k,n}^*, \nu_{k,n}^*),$$
$$\hat{\alpha} = \sum_k \hat{\boldsymbol{\alpha}}_k, \quad \nu_{k,n}^* = \hat{\nu}_k + 1,$$
$$W_{k,n}^{*\,-1} = \hat{W}_k^{-1} + (\mathbf{x}^* - \mathbf{x}_n)(\mathbf{x}^* - \mathbf{x}_n)^T. \tag{5}$$

It is noteworthy that the computation of the matrix parameter $W_{k,n}^{*\,-1}$ in Eq. 5 depends not only on the matrix parameter $\hat{W}_k^{-1}$ from Wishart distribution of trained metric bases but also the relationship between training instance $\mathbf{x}_n$ and test instance $\mathbf{x}^*$. To simplify the computation in Eq. 5, we can use Woodbury Matrix Identity to get an equivalent update form: $W_{k,n}^* = \hat{W}_k - \hat{W}_k(\mathbf{x}^* - \mathbf{x}_n)(1 + (\mathbf{x}^* - \mathbf{x}_n)^\top\hat{W}_k(\mathbf{x}^* - \mathbf{x}_n))^{-1}(\mathbf{x}^* - \mathbf{x}_n)^\top\hat{W}_k$. The distance from $\mathbf{x}^*$ to a training

instance $\mathbf{x}_n$ is $(\mathbf{x}^* - \mathbf{x}_n)^\top \mathbb{E}[M_{\mathbf{x}^*}](\mathbf{x}^* - \mathbf{x}_n)$. The term with most cost substantially relies on the repeated computation of distance between new instance $\mathbf{x}^*$ and all the training data, which is measure by $\mathbb{E}[M_k]$, i.e., the expectation of distribution over metric bases. The overall complexity of computing the nearest neighbor for $\mathbf{x}^*$, consequently, is $O(NKd^2)$. With parallel computing, it can be reduced to $O(d^2)$ by inferring each instance/local basis separately.

It is remarkable that the posterior expectation over specific metric for an *unseen* instance is a combination of information from both trained metric bases and the posterior of $\mathbf{z}$. This agrees with the previous statement that ISMETS learns metric subspace $\mathcal{M}_{\mathbf{x}^*}$ spanned by metric bases. More importantly, $\mathbf{z}$ is drawn from a Dirichlet prior and can be sparse, therefore limited metric bases are substantially used in consisting of the $M_{\mathbf{x}}$. In other words, each ISM is sampled from a subspace $\mathcal{M}_{\mathbf{x}} \subseteq \mathcal{M}$, and thus contains local properties of data. Eq. 5 also indicates that the posterior $p(M_{\mathbf{x}^*})$ is related to the distances from $\mathbf{x}^*$ to all others, so $M_{\mathbf{x}^*}$ can imply some global information of the metric space as well.

Experiments

In this section, we first illustrate the mechanism of ISMETS with synthetic data, and then compare ISMETS with other distance metric learning methods. In order to investigate the abilities of handling unlabeled data and the robustness of ISMETS, we conduct more experiments to discover the influence on label ratio and the number of metric bases.

Experiments on Synthetic Data

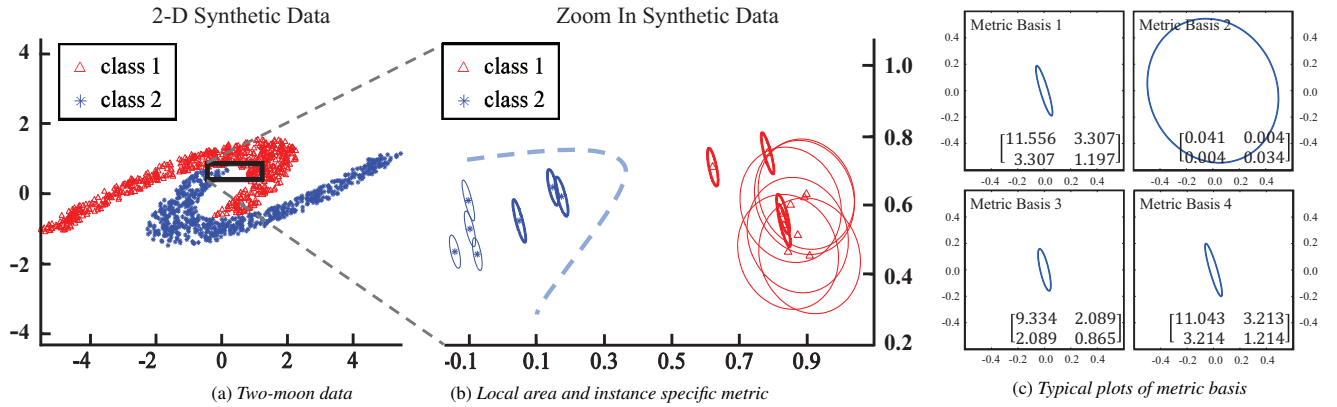
A two-moon synthetic data with two classes (Haykin 2009) is used to show that the learned results of ISMETS are with good *interpretability*. Each class contains 1000 instances. We randomly choose 70% of the data as labeled and the remains are unlabeled. The number of metric bases is set as 20. Only labeled synthetic data is plotted in Fig. 2 (a). ISMETS first learns all metric bases, and in Fig. 2 (c), 4 typical plots of metric basis are plotted. The size and rotation of each ellipsoid depend on the inverse of a metric basis. Values of bases are shown in the right-bottom of each subplot of (c) and all ellipsoids are centered at $(0,\ 0)$. It can be clearly found that these 4 bases are different. The specific metric $M_{\mathbf{x}}$ for each instance $\mathbf{x}$ in a small part of the two-moon data (in Fig. 2 (a), marked with black rectangle) is plotted in Fig. 2 (b); each ellipsoid in (b) depends on the inverse of the concerned instance specific metric, and its center is located on the corresponding instance. The blue dot-curve in (b) reflects part of the decision boundary in local area. The rotation directions of metrics for instances from both classes are obviously consistent with the decision boundary. Instances in each class near the blue dot-curve are with smaller ellipsoids and are marked in bold, while those instances far away from the dot-curve are with larger ellipsoids. This indicates the points around the boundary are with larger specific metrics. Therefore, ISMETS trends to push instances around the boundary farther and makes lower density area around boundary for better classification performance.

Experiments on Real Datasets

We test ISMETS on 8 UCI datasets and 4 real Bioinformatic datasets[1] (GDS3286, GSE4115, GDS2771, GDS531). 3 of 8 UCI datasets are multi-class tasks and marked with "*" in Table 1. ISMETS is compared with 8 state-of-the-art metric learning methods, i.e., SCML_{G}, SCML_{L} (Shi, Bellet, and Sha 2014), ISD-L1, ISD-L2 (Zhan et al. 2009), LMNN, mmLMNN (Weinberger and Saul 2009), DNE (Zhang et al. 2007), SDA (Cai, He, and Han 2007). Among compared methods, $\text{SCML}_{\text{G/L}}$ refer to the global and local sparse compositional metric learner respectively, ISD-L1/L2 are methods with different losses, and mmLMNN is a multiple metric version of LMNN. kNN with Euclidean distance is also listed as Euclid in Table 1. As researchers did in almost all distance metric learning literatures, all compared methods invoke kNN as the classifier and then use classification errors for evaluating the quality of learned metrics to compare with each other. In our implementation, k is configured as 5. We run each method 30 random trials per data. The mean and standard derivation of classification errors are listed in Table 1. In each trial, we randomly split the data into training set (67%) and test set (33%). In the training set, 30% data are labeled examples and the remains are unlabeled. For Bioinformatic data, PCA is employed for projecting the data into a low dimensional feature space with the dimensionality equals the number of instances. Since ISMETS can predict both transductively and inductively, we denote the results as ISMETS_{t} and ISMETS_{i} respectively in following tables and figures. The maximum training iteration for $\text{ISMETS}_{\text{t/i}}$ is fixed as 20. We use non-informative hyper-parameters in the training process (Gönen and Margolin 2014; Zhao et al. 2014), i.e., we set α_σ, β_σ and elements in $\boldsymbol{\alpha}_\pi$ all as 1.

In Table 1, the last two rows give the Win/Tie/Lose counts of t-test results at a significant level 95% for $\text{ISMETS}_{\text{t/i}}$ vs. others. From the results, it can be found that $\text{ISMETS}_{\text{t/i}}$ achieve the best performance among all methods on 6 datasets, while $\text{SCML}_{\text{G/L}}$ outperform others on 3 datasets, ISD-L1/L2 achieve the best on 2, LMNN series and SDA are superior to others on 1 dataset respectively. According to t-test results, it reveals that both $\text{ISMETS}_{\text{t/i}}$ can yield better performance than compared methods. In particular, compared with ISD-L1/L2, mmLMNN, DNE, SDA and Euclid, ISMETS_{t} never loses and similarly, ISMETS_{i} never loses to ISD-L1/L2, mmLMNN, DNE and Euclid according to t-test at 95% significance level. We have marked the corresponding 0 loss results with "*" in Table 1, and the wins number of $\text{ISMETS}_{\text{t/i}}$ are bolded. Results in Table 1 reveals the *effectiveness* of ISMETS. The probable reason for the better performance of ISMETS to the global methods may be the consideration of the local structure and its superiority to the local ones may be due to the simultaneous learned bases/coefficients in the model. Moreover, experiments running on computational servers with 2.66GHz 2 cores and 4GB memory show that ISMETS can be trained faster than some other ISM type methods. E.g., on mfeat-mor dataset, ISMETS_{i} is on average 7.82 and 1.72 times faster than ISD-L1 and mmLMNN respectively.

[1] Can be open accessed from www.ncbi.nlm.nih.gov/geo/



(a) *Two-moon data* (b) *Local area and instance specific metric* (c) *Typical plots of metric basis*

Figure 2: Interpretability illustrations of ISMETS on *two-moon* data. Subplot (a) displays the data distribution together with a selected area marked in black rectangle. Subplot (b) zooms in the selected area of (a) and marks the instance specific metrics learned by ISMETS. Subplot (c) shows 4 of learned metric bases. All ellipsoids are drawn according to the inverse of metrics.

Table 1: Comparisons of classification performance (test errors, mean ± std.) with other methods. ISMETS_t represents ISMETS with transductive strategy while ISMETS_i predicts in an inductive manner. Last two rows list the Win/Tie/Lose counts on all datasets with t-test against other methods at significance level 95%. The win counts for $\text{ISMETS}_{\text{t/i}}$ to other compared methods and the best performance on each dataset are bolded. The W/T/L is marked with "*" where $\text{ISMETS}_{\text{t/i}}$ never lose.

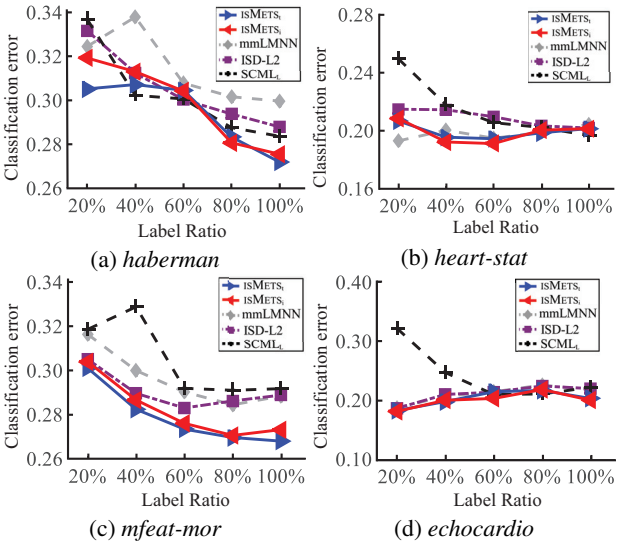
Name	ISMETS_t	ISMETS_i	SCML_G	SCML_L	ISD-L1	ISD-L2	mmLMNN	LMNN	DNE	SDA	Euclid
*balance**	.161±.021	.162±.026	**.117±.032**	.121±.032	.175±.021	.168±.023	.239±.036	.170±.017	.174±.021	.271±.088	.170±.015
echocardio	**.191±.031**	.193±.031	.261±.059	.264±.074	.196±.054	.203±.058	.208±.050	.207±.049	.208±.047	.195±.029	.208±.050
haberman	.306±.046	.315±.046	**.296±.033**	**.296±.033**	.303±.037	.316±.048	.321±.040	.314±.044	.301±.037	.318±.043	.303±.038
heart-stat	.196±.038	.193±.038	.200±.119	.233±.119	**.191±.033**	.216±.037	.197±.033	.195±.035	.197±.033	.407±.090	.197±.033
liver-dis	.404±.048	**.400±.040**	.426±.059	.426±.059	.418±.049	.412±.049	.425±.034	.406±.052	.409±.053	.450±.059	.409±.053
*mfeat-mor**	.285±.018	.287±.018	.286±.018	.317±.016	.295±.016	.295±.015	.307±.019	.296±.014	.299±.016	**.281±.015**	.293±.016
*page-blo**	**.052±.005**	**.052±.005**	.059±.010	.063±.014	.063±.005	.067±.004	.064±.005	.060±.004	.060±.004	.055±.006	.059±.004
vote	**.066±.049**	.080±.049	.142±.137	.153±.141	.082±.032	.074±.030	.154±.000	.085±.032	.088±.029	.371±.097	.088±.031
GDS3286	.338±.000	**.336±.005**	.355±.038	.355±.038	.472±.145	.404±.114	.409±.089	.432±.116	.441±.114	.526±.176	.441±.114
GSE4115	.469±.019	.464±.026	.485±.078	.489±.081	.487±.035	**.461±.028**	.481±.050	.472±.030	.479±.034	.481±.015	.479±.034
GDS2771	.475±.032	.469±.017	.487±.058	.476±.046	.480±.014	.475±.026	**.462±.044**	.476±.026	.477±.022	.479±.013	.477±.022
GDS531	**.207±.000**	**.207±.000**	**.207±.000**	**.207±.000**	.320±.150	.280±.095	.235±.044	.330±.118	.348±.140	.243±.138	.348±.140
W / T / L	ISMETS_t vs. others		**5** / 6 / 1	**6** / 5 / 1	**6** / 6 / 0 *	**6** / 6 / 0 *	**7** / 5 / 0 *	**6** / 6 / 0 *	**6** / 6 / 0 *	**7** / 5 / 0 *	**6** / 6 / 0 *
W / T / L	ISMETS_i vs. others		**5** / 6 / 1	**6** / 5 / 1	**7** / 5 / 0 *	**6** / 6 / 0 *	**7** / 5 / 0 *	**6** / 6 / 0 *	**6** / 6 / 0 *	**8** / 3 / 1	**5** / 7 / 0 *

Fig. 3 shows the influence over classification vs. label ratio, i.e., the portion of labeled data in the training set, changing from 20% to 100% on 4 datasets. The trends on average errors of *instances specific or local* distance metric methods are reported. The average errors of compared methods are decreased on most datasets when more labeled examples are used, while $\text{ISMETS}_{\text{t/i}}$ perform better than others in most cases. Especially, when the label ratio is 20%, $\text{ISMETS}_{\text{t/i}}$ almost outperform other methods. This indicates $\text{ISMETS}_{\text{t/i}}$ can make full use of the information from unlabeled data and adapt to semi-supervised scenarios when labeled examples are very limited.

To investigate the influence on the number of metric bases, we conduct more experiments on $\text{ISMETS}_{\text{t/i}}$ with different choice of metric bases number K. The changes of mean classification error are plotted in Fig. 4. We can find that the errors of $\text{ISMETS}_{\text{t/i}}$ only vary in a small range when the number of metric bases changes. Thus, tuning the number of bases may not impact performance dramatically, i.e., ISMETS is *robust* to the number of metric bases.

Conclusion

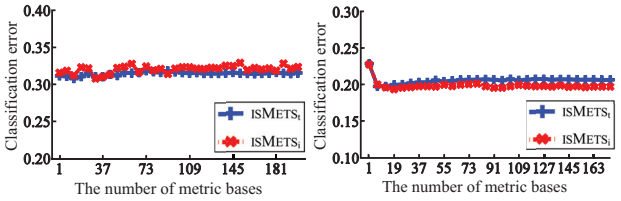
We propose a unified framework ISMETS to tackle instance specific metric learning where we can predict the specific metrics for *unseen* test instances *inductively* as well as *transductively*. In transductive setting, the metrics of instances are yielded by an expectation over a latent allocation distribution, while the instances' metrics can be obtained by integrating out the latent allocation variables in inductive setting. In addition, ISMETS automatically learns metric bases and form instance specific subspaces with sparse combination of those bases. *Interpretability* of our proposed method is showed on synthetic data. Moreover, extensive evaluations have demonstrated its *effectiveness* and *robustness*. Note that



(a) *haberman* (b) *heart-stat*

(c) *mfeat-mor* (d) *echocardio*

Figure 3: Influence on label ratio.



(a) *haberman* (b) *heart-stat*

Figure 4: Influence on the number of metric bases.

the supervision part of the model can be substituted with other forms, which implies ISMETS framework is general for instance specific metric learning. For faster metric learning and prediction, we will try to incorporate parallel techniques and some approximation tricks into our framework.

References

Babagholami-Mohamadabadi, B.; Roostaiyan, S. M.; Zarghami, A.; and Baghshah, M. S. 2014. Multi-modal distance metric learning: A bayesian non-parametric approach. In *Proceedings of the 13th European Conference on Computer Vision Workshops*, 63–77.

Bellet, A.; Habrard, A.; and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cai, D.; He, X.; and Han, J. 2007. Semi-supervised discriminant analysis. In *Proceedings of the 11th International Conference on Computer Vision*, 1–7.

Chen, W.; Chen, Y.; and Weinberger, K. Q. 2014. Fast flux discriminant for large-scale sparse nonlinear classification. In *Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining*, 621–630.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning*, 209–216.

Frome, A.; Singer, Y.; and Malik, J. 2007. Image retrieval and classification using local distance functions. In *Advances in Neural Information Processing Systems 19*. Cambridge, MA.: MIT Press. 417–424.

Gönen, M., and Margolin, A. A. 2014. Kernelized bayesian transfer learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 1831–1839.

Haykin, S. O. 2009. *Neural networks and learning machines*. Upper Saddle River, NJ.: Prentice Hall.

Hu, J.; Zhan, D.-C.; Wu, X.; Jiang, Y.; and Zhou, Z.-H. 2015. Pairwised specific distance learning from physical linkages. *ACM Transactions on Knowledge Discovery from Data* 9(3):Article 20.

Jin, R.; Wang, S.; and Zhou, Y. 2010. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems 23*. Cambridge, MA.: MIT Press. 862–870.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning* 37(2):183–233.

Kulis, B. 2012. Metric learning: A survey. *Foundations and Trends in Machine Learning* 5(4):287–364.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems 20*. Cambridge, MA.: MIT Press. 121–128.

Shi, Y.; Bellet, A.; and Sha, F. 2014. Sparse compositional metric learning. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2078–2084.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1(1-2):1–305.

Watanabe, K., and Watanabe, S. 2006. Stochastic complexities of gaussian mixtures in variational bayesian approximation. *Journal of Machine Learning Research* 7:625–644.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10:207–244.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18*. Cambridge, MA.: MIT Press: MIT Press. 1473–1480.

Xiang, S.; Nie, F.; Meng, G.; Pan, C.; and Zhang, C. 2012. Discriminative least squares regression for multiclass classification and feature selection. *IEEE Transactions on Neural Networks and Learning Systems* 23(11):1738–1754.

Xing, E. P.; Ng, A. Y.; Jordan, M. I.; and Russell, S. 2003. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*. Cambridge, MA.: MIT Press. 505–512.

Yang, L.; Jin, R.; and Sukthankar, R. 2007. Bayesian active distance metric learning. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, 442–449.

Zhan, D.-C.; Li, M.; Li, Y.-F.; and Zhou, Z.-H. 2009. Learning instance specific distances using metric propagation. In *Proceedings of the 26th International Conference on Machine Learning*, 1225–1232.

Zhang, W.; Xue, X.; Sun, Z.; Guo, Y.-F.; and Lu, H. 2007. Optimal dimensionality of metric space for classification. In *Proceedings of the 24th international conference on Machine learning*, 1135–1142.

Zhao, Q.; Meng, D.; Xu, Z.; Zuo, W.; and Zhang, L. 2014. Robust principal component analysis with complex noise. In *Proceedings of the 31st International Conference on Machine Learning*, 55–63.